TECHNICAL RESEARCH REPORT

Opportunistic packet scheduling in cellular networks with base station antenna arrays

by Tianmin Ren, Richard J. La

CSHCN TR 2006-2
(ISR TR 2006-4)

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Opportunistic Packet Scheduling in Cellular Networks with Base Station Antenna Arrays	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland, College Park,Department of Electrical & Computer Engineering,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We study the issue of designing a downlink scheduling policy for a cellular network with base station antenna arrays. We derive an optimal scheduling policy that achieves the throughput region which is a set of feasible arrival rate vectors that can be stabilized by some scheduling policy. Then based on the structure of the derived optimal policy whose complexity increases exponentially with the number of users in the system, we propose two heuristic scheduling algorithms with much lower complexity. We demonstrate that our proposed algorithms perform much better than other heuristic algorithms that do not take into consideration the physical layer constraints and/or queue lengths in the sense that they have a larger throughput region than other heuristic algorithms.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**39**	

Opportunistic Packet Scheduling in Cellular Networks with Base Station Antenna Arrays

Tianmin Ren and Richard J. La

Department of Electrical & Computer Engineering

University of Maryland, College Park

{rtm,hyongla}@eng.umd.edu

Abstract

We study the issue of designing a downlink scheduling policy for a cellular network with base station antenna arrays. We derive an optimal scheduling policy that achieves the throughput region, which is a set of feasible arrival rate vectors that can be stabilized by some scheduling policy. Then, based on the structure of the derived optimal policy whose complexity increases exponentially with the number of users in the system, we propose two heuristic scheduling algorithms with much lower complexity. We demonstrate that our proposed algorithms perform much better than other heuristic algorithms that do not take into consideration the physical layer constraints and/or queue lengths in the sense that they have a larger throughput region than other heuristic algorithms.

Keywords – antenna array, cross-layer design, optimal scheduling, stability

I. Introduction

Wireless communication has been going through a rapid transition from the traditional circuit switched voice services to packet switched data services during the past decade. Increasing demand for fast wireless access and high-speed wireless links and the advent of wireless applications such as wireless multimedia that demands certain quality-of-service (QoS) guarantees, have stimulated much research in wireless communication.

New network architectures and protocols are proposed to support data applications in wireless networks. A typical architecture in many of current wireless systems, especially cellular networks (*e.g.,* UMTS, cdma 2000), provides wireless access to mobile users through access points (APs) or base stations (BSs) that are connected to the core network. The most challenging task in designing these communication systems is to support QoS guarantees to various data applications on wireless channels with limited bandwidth and time varying characteristics. There exist different notions of QoS at different communication layers. For instance, the QoS at the physical layer may be given as certain signal to interference and noise ratio (SINR) or a corresponding bit error rate (BER) at the receiver. At the MAC layer the QoS is typically expressed in terms of achievable bit rate or packet error rate (PER), while at the higher layers the QoS can be perceived as a minimum throughput and/or maximum delay requirement. The ability of a network to satisfy such QoS requirements and enhance the system capacity depends on the interaction/cooperation of several layers. As a result, a joint design of protocols for two or more layers can lead to significant improvement in overall network performance. This paper provides an example where a joint design of physical layer and MAC scheduling algorithms offers a significant improvement in network throughput/capacity.

A wide spectrum of approaches have been proposed to reuse the communication resources in time, frequency and/or space domain, to provide the QoS guarantees to mobile users and to improve wireless network capacity. Among these approaches, the application of antenna arrays, which exploits the spatial diversity of mobile users, is considered a more promising one and the last frontier for future capacity improvement of wireless networks. This is due to the beamforming capability of the antenna arrays that can form the beam pattern directed to a desired user while nulling out the others. This greatly reduces co-channel interference at the receiver(s), and spatially separable users can share the same channel with their QoS requirements satisfied.

Previous research on application of antenna arrays in cellular environments can be categorized into two classes. The first class of research is on the physical layer; given a set of users, the

problem is to design optimal algorithms to calculate the beamforming weights for each user. The problem is often modeled as an optimization problem, where the objective is to minimize the total transmission power subject to the minimum SINR constraint for each user [3], [14].

The second class of research is on the MAC layer with consideration of physical layer user separability constraints. The goal of this approach is, given a set of users, to place as many users on the same channel as possible and compute the beamforming weights for each selected user subject to the SINR constraint. This helps maximize the (instant) throughput of the network. This problem is also studied with other multi-user access schemes such as TDMA, OFDM and CDMA [5]. All of these studies, however, assume that users are modeled as infinite sources where there is always a packet to be served at the queue for each user. Major drawbacks of these works are the limitation of the focus on instant total throughput maximization and the lack of consideration of upper layer QoS requirements for individual users. Thus, the assignment of users on a channel only reflects the feasibility at the physical layer, but not the current buffer occupancy or traffic demand of each user. This separation of physical layer algorithms and upper layer QoS requirements leads to degradation of overall system performance.

In this paper we study the problem of designing a scheduling algorithm with BSs that are equipped with an antenna array. We first consider the case where a central controller handles multiple BSs serving a set of users. In this case packets arrive at the central controller for transmission to mobile users. We model the system as a queueing system with multiple parallel servers, and the physical layer constraints are imposed on the selection of users that can be served in each timeslot. Instead of a policy that maximizes instant throughput, we identify an optimal scheduling policy that stabilizes the system if it is stable under some policy. The user throughput requirements are satisfied under this optimal scheduling policy and, hence, the long term total system throughput is maximized.

Similar queueing systems have been used to model other scenarios in [1], [6], [10], [19], [20], and were first proposed in [19] for a multi-hop radio network. They derive an optimal policy

of similar form (based on maximum weighted queue size). However, the complexity of these optimal scheduling policies increases exponentially with the number of users/queues, and no practical sub-optimal scheduling policy is proposed. In this paper, we follow a similar approach as in [10], and propose two scheduling policies with significantly lower complexity that achieve sub-optimal performance for our problem. In fact, the first proposed scheduling algorithm has linear complexity.

Some of our preliminary work is presented in [15]. However, the performance evaluation reported in [15] was carried out using a simple heuristic beamforming algorithm, assuming all transmissions are successful. In this paper we use an optimal beamforming algorithm proposed in [17] and explicitly model the events of successful transmissions based on the achieved SINR values at the scheduled users and link curves. Most importantly, in [15] we consider a multiple cell network under centralized control and do not study the performance of our algorithms in multiple cell networks where resource allocations are carried out by the BSs in a distributed manner without coordination. In such an environment the inter-cell interference caused by other co-channel cells at a scheduled user is *unknown* in advance. As a result the traditional approach of aiming to achieve some target SINR at the scheduled users, which is the same approach taken for our single cell network study, is no longer feasible.

Understanding the effects of the unknown inter-cell interference on the performance of the beamforming and scheduling algorithms is crucial as many of the wireless service providers are currently interested in moving the intelligence from a central resource manager (*e.g.,* radio network controller (RNC)) to the edge of the network (*e.g.,* BSs) in order to handle the increasing complexity of the resource allocation algorithms and to reduce the delays in channel estimation between the BSs and the central resource manager. In this paper we adopt a beamforming algorithm we propose in [16] (summarized in subsection VI-B in this paper). This algorithm is a modification of the beamforming algorithm proposed in [17] and provides guaranteed average PERs to the users in the presence of unknown inter-cell interference. Using this beamforming

algorithm we investigate the performance of our proposed scheduling algorithms and the degradation in network performance due to the unknown inter-cell interference in Section VII. We also carry out a study on the effects of the inter-cell interference on network performance and spectral efficiency under different reuse patterns.

This paper is organized as follows. In Section II we describe the problem of designing an efficient downlink scheduling algorithm with base station antenna arrays, and derive an optimal scheduling policy based on feasible rate vectors. In Section III we describe the single cell network model, which is followed by our proposed heuristic algorithms that approximate the optimal scheduling policy with lower complexity in Section IV. Simulation results of the proposed algorithms in a single cell network are given in Section V. Section VI describes a multi-cell network model and beamforming algorithms that handle *random* inter-cell interference in such environments, followed by simulation results of the proposed algorithms in a multi-cell environment in Section VII. We conclude in Section VIII.

II. Optimal Downlink Scheduling

In this section we consider the case where a centralized control agent carries out scheduling and beamforming for multiple BSs serving a set of users. We define an achievable rate vector and a throughput region, and present an optimal scheduling algorithm that can achieve any interior point in the throughput region.

A. System model

We consider a wireless network that consists of several BSs. Each BS is equipped with an antenna array so that several users can be served simultaneously. These BSs are coordinated by a single central controller. Mobile users in the network are able to receive data packets from any of these BSs. However, at any given time, a mobile user can receive data packet(s) from at most one BS. We assume a time slotted system where the transmission time of a packet equals the duration of a timeslot when the lowest transmission rate is selected. In each timeslot, the

central controller collects the information regarding the wireless channel conditions of each user to different BSs. Based on this information and the number of backlogged packets of each user, the central controller makes a scheduling decision for the timeslot. The scheduling decision made by the central controller includes assignment of BSs to the users and the transmission rate of each user, and the calculation of the beamforming weights for the selected transmission rates.

The system architecture under study is depicted in Fig. 1. User packets enter the scheduling module at the central controller, which determines the assignments of BSs and transmission rates. Beamforming and power adaptation are subsequently carried out for scheduled users. Scheduling and beamforming are *interdependent* operations that depend on network state (*i.e.*, queue sizes) and channel state information, which is assumed to be available at the central controller.

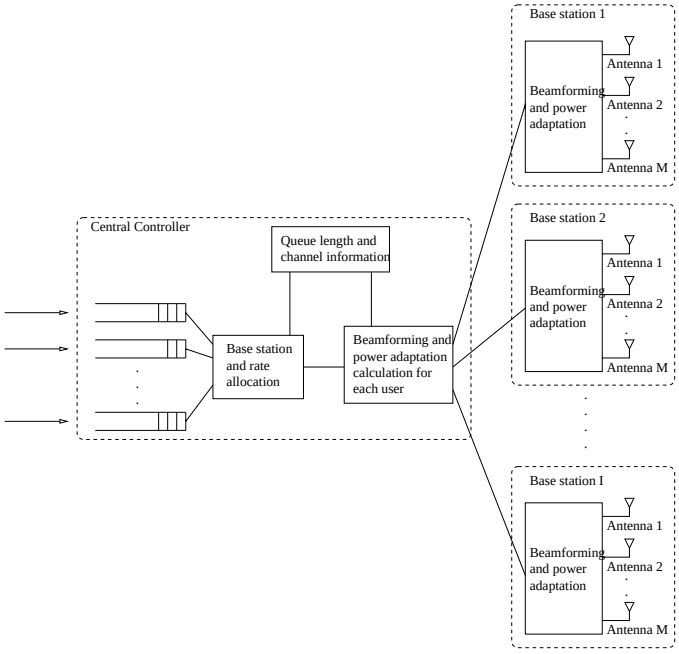


Fig. 1. The multiple cellular communication system.

B. Problem statement

The network consists of I BSs shared by J mobile users. We denote the set of BSs by $\mathcal{I} = \{1, \ldots, I\}$ and the set of users by $\mathcal{J} = \{1, \ldots, J\}$. There is a central controller that

coordinates the operation of the I BSs. Each BS is equipped with an M-element antenna array. Several transmission rates are available at the BSs based on the channel conditions. The set of available transmission rates is denoted by $\mathcal{V}$. We assume that each transmission rate is a positive integer number. If rate $v \in \mathcal{V}$ is chosen, up to v packets can be transmitted in one timeslot, depending on the number of packets waiting for transmission. We denote $|\mathcal{V}| = V$.

Packets arrive at the central controller for transmission. The central controller maintains a separate queue for each user. Let $a_j(t), j = 1, 2, \cdots, J$ and $t = 0, 1, \cdots$, denote the number of packets that arrive at queue j in timeslot t. We assume that $a_j(t), t = 0, 1, 2, \ldots,$ are given by independent and identically distributed (i.i.d.) random variables (rvs) with a finite second moment, *i.e.,* $\mathbf{E}\left[a_j(t)^2\right] < \infty$.[1] The average arrival rate of user j, $\mathbf{E}\left[a_j(t)\right]$, is denoted by A_j. We call the vector of the average arrival rates of the users, $\mathbf{A} = (A_1, A_2, \cdots, A_J)^T$, an arrival vector.

We assume that the central controller has perfect channel information of each user to the BSs. In each timeslot, the central controller (i) assigns the BSs to the users, (ii) computes transmission rates of scheduled users, and (iii) calculates the beamforming weights of the scheduled users. A scheduling decision by the central controller can be expressed as an $I \times J$ matrix $\mathbf{R} = [r_{ij}]$ where the element $r_{ij} \in \mathcal{V} \cup \{0\}, i = 1, \cdots, I$, and $j = 1, \cdots, J$, is the transmission rate of BS i to user j. However, a rate matrix $\mathbf{R}$ can be selected for transmission by the central controller only if it satisfies certain physical layer constraints described below.

For data traffic, in order to maintain reasonable performance at higher layers the PER at the physical/link layer needs to be kept fairly lower. This PER requirement demands that the SINR at each receiver be above some threshold value. A rate matrix is feasible if and only if (i) each user receives packets from at most one BS and (ii) SINR requirement is satisfied for every user. Note that the feasibility of a rate matrix depends on the target PER that determines the SINR

[1]The results in this paper can be easily extended to the case where the arrival processes are given by an ergodic Markov chain.

requirement for each user.

We model the channel process for all users as a finite state ergodic Markov chain (MC) with a stationary distribution π. Each channel state is associated with a set of feasible rate matrices. In other words, a state of the MC specifies the set of all feasible rate matrices that can be selected for transmission given the channel conditions of the users. The state space of the MC is denoted by $\mathcal{S}$. The problem we are interested in is one of finding a scheduling policy that selects a feasible rate matrix given the queue sizes and channel state in each timeslot, so that the system is stable whenever possible under some policy. In this paper we only consider stationary policies, *i.e.,* the scheduling decisions do not depend on timeslot t, but only on the channel state $\mathbf{S}(t)$ and the queue sizes $\mathbf{X}(t) := (x_1(t), \ldots, x_J(t))^T$, where $x_j(t)$ is the number of backlogged packets in queue j at the beginning of timeslot t. A stationary scheduling policy can be viewed as a mapping that assigns to each pair $(\mathbf{X}, \mathbf{S}) \in \mathbb{Z}_+^J \times \mathcal{S}$ of queue sizes and channel state a probability distribution on the set of feasible rate matrices for the given state $\mathbf{S}$, where $\mathbb{Z}_+ := \{0, 1, \ldots\}$, where the probability assigned to a feasible rate matrix is the probability the policy will select the rate matrix for transmission given $\mathbf{S}(t)$ and $\mathbf{X}(t)$.

C. Throughput region

In this subsection we first define a stable arrival vector and then characterize the throughput region.

Definition 1: An arrival vector $\mathbf{A}$ is said to be stable if there exists a scheduling policy such that

$$\lim_{c\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}_{[x_j(\tau)>c]} = 0 \ , \qquad \text{for all } j = 1, 2, \cdots, J \ . \tag{1}$$

If a scheduling algorithm satisfies (1), then we say that $\mathbf{A}$ is stable under the scheduling policy. The throughput region, denoted by $\mathcal{A}$, is defined to be the closure of the set of stable arrival vectors.

The following proposition characterizes the throughput region $\mathcal{A}$.

Proposition 1: A necessary and sufficient condition for an arrival vector $\mathbf{A}$ to belong to $\mathcal{A}$, is that there exists a scheduling policy that achieves

$$\mathbf{A} \leq \mathbf{D} := \sum_{\mathbf{S} \in \mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R} \in \mathbf{S}} c_{\mathbf{SR}} \mathbf{R}^T \mathbf{1}_{I \times 1} (1 - PER) \tag{2}$$

where $c_{\mathbf{SR}}, \mathbf{S} \in \mathcal{S}, \mathbf{R} \in \mathbf{S}$, are nonnegative numbers such that $\sum_{\mathbf{R} \in \mathbf{S}} c_{\mathbf{SR}} = 1$ for all $\mathbf{S} \in \mathcal{S}$.

Proof: A proof is provided in Appendix I ■

D. Optimal scheduling policy

In this subsection we describe an optimal scheduling policy that satisfies (1) for every $\mathbf{A} \in int(\mathcal{A})$, where $int(\mathcal{A})$ denotes the interior of the throughput region $\mathcal{A}$. In particular, we consider the following scheduling policy: Given backlog vector $\mathbf{X}(t)$ and channel state $\mathbf{S}(t)$, the rate matrix selected by the scheduling algorithm is given by

$$\mathbf{R}(t) = \arg \max_{\mathbf{R} \in \mathbf{S}(t)} \mathbf{X}(t)^T (\mathbf{R^T}\ \mathbf{1}_{I \times 1}) \ . \tag{3}$$

Ties are assumed to be broken arbitrarily.

The following proposition establishes the (throughput) optimality of the scheduling policy given by (3).

Proposition 2: Suppose that $\mathbf{A} \in int(\mathcal{A})$. Then, the system is stable under the scheduling policy given by (3).

Proof: A proof is given in Appendix II. ■

In fact, in the proof we prove that any stationary scheduling policy that selects the rate matrix $\arg\max_{\mathbf{R} \in \mathbf{S}(t)} \mathbf{X}(t)^T \Xi (\mathbf{R^T}\ \mathbf{1}_{I \times 1})$, where $\Xi = \text{diag}(\xi_j, j \in \mathcal{J})$ and $\xi_j > 0$, is a throughput optimal policy and the system is stable if the arrival vector $\mathbf{A} \in int(\mathcal{A})$.

In this section we have considered scenarios where a centralized controller carries out the scheduling and beamforming for multiple BSs serving a fixed set of users, and derived an optimal scheduling policy. The derived optimal scheduling policy, however, does not yield a

practical implementation as it requires searching through *all* feasible rate matrices given the current channel state and identifying the one that maximizes the inner product given in (3).

A natural question that arises is how one can design more practical scheduling algorithms based on the optimal scheduling policy. In the following sections, we assume that each user is associated with the closest BS, *i.e.,* static BS assignment, and investigate the issue of designing practical scheduling algorithms with an antenna array at the BS(s). We will first consider a simple case of single cell networks in Sections III and V, and then discuss multiple-cell environments in Sections VI and VII. In the case of a multiple-cell network, a receiving user experiences inter-cell interference from co-channel cell BSs that share the same frequency spectrum. Therefore, a BS needs to compensate for this random inter-cell interference experienced by a receiver when computing beamforming weights and transmission power of a scheduled user. This issue will be discussed in Section VI.

III. Single Cell Network Model

In this section we first describe the channel models used in our study and then present the optimal beamforming algorithms to be used in our study of single cell networks.

A. Single cell channel model

In this subsection we describe the wireless channel model for a single cell network with a cell radius of R. A BS is located at the center of the cell and transmits packets to N users uniformly distributed in the cell. In each timeslot, the BS schedules a set of users for transmission, and calculates their beamforming weights and transmission powers.

The BS is equipped with an antenna array, where M antenna elements are uniformly located on a circle of radius r. The multi-path channel between a given user and the m-th antenna element of the BS is expressed as $h^m(t) = \sum_{\ell=1}^{L} g_\ell \, \delta\left(t - \tau_\ell - \tau_\ell^m\right)$, where L is the number of paths, g_ℓ is the complex gain of the ℓ-th path, and τ_ℓ is the delay of the ℓ-th path with respect to the first antenna element with $m = 1$. The gain g_ℓ is a complex rv with zero mean and

variance A_ℓ. The term $\tau_\ell^m = (r/c)(\cos\theta_\ell - \cos(2\pi(m-1)/M - \theta_\ell))$ captures the delay to the m-th antenna with respect to the first antenna, and θ_ℓ is the angle of arrival of the ℓ-th path of the user, and c is the electro-magnetic wave propagation speed. We assume each path results from the reflection by a scatterer, and L scatterers are distributed within a circle of radius R' centered at the user and are uniformly distributed both in distance and in angle with reference to the user.

The signal received by the user from the BS is given by

$$y(t) = \sqrt{p} \sum_{m=1}^{M} w^m \sum_{\ell=1}^{L} g_\ell \; e^{j\omega\tau_\ell^m} s(t-\tau_\ell) \;,$$

where p is the transmission power, w^m is the beamforming weight of the m-th antenna element, and $s(t)$ is the signal. Note that we implicitly assume that $\tau_\ell \gg \tau_\ell^m$ and the sum $\tau_\ell + \tau_\ell^m \approx \tau_\ell$. Beamforming vector $\mathbf{w} = [w^1, w^2, \cdots, w^M]^T$ satisfies $\mathbf{w}^H\mathbf{w} = 1$. The $M \times 1$ antenna steering vector $\mathbf{v}(\theta_\ell)$ at direction θ_ℓ is defined to be $[e^{j\omega\tau_\ell^m};\; m = 1, \ldots, M]$, where ω is the carrier frequency. If we assume that all paths are independent, the expected received signal power is given by

$$\mathbf{E}\left[\left|\sqrt{p} \sum_{m=1}^{M} w^m \sum_{\ell=1}^{L} g_\ell \; e^{j\omega\tau_\ell^m} s(t-\tau_\ell)\right|^2\right] = p\mathbf{w}^H\mathcal{H}\mathbf{w} \;,$$

where $\mathcal{H} = \sum_{\ell=1}^{L} A_\ell \mathbf{v}(\theta_\ell)\mathbf{v}^H(\theta_\ell)$. This is because $\mathbf{E}\left[g_{\ell_1} g_{\ell_2}^*\right]$ equals A_ℓ if $\ell_1 = \ell_2$ and 0 otherwise from the assumption $\mathbf{E}[g_\ell] = 0$. The matrix $\mathcal{H}$ is called a spatial covariance matrix and in general has a rank larger than one.

For time varying channels, the variance of the channel gain $\{A_\ell(t); t = 1, 2, \ldots\}$ is a stochastic process. We consider *temporally correlated shadow fading plus Rayleigh fading channel model* in this paper. The rv $A_\ell(t) = (s_\ell(t) f_\ell(t))^2 / d_\ell^\kappa$, where $s_\ell(t)$ and $f_\ell(t)$ are sequences of log-normal and Rayleigh rvs, respectively, accounting for slow shadow fading and fast fading. The variable d_ℓ denotes the distance from the BS to the user along the ℓ-th path and κ is the path loss exponent. The sequence of rvs $\{s_\ell(t)\} = \{e^{r_\ell(t)}\}$, where $\{r_\ell(t)\}$ is a sequence of Gaussian rvs

generated from $r_\ell(t+1) = (1-\rho)r_\ell(t) + \rho \cdot u_\ell(t)$ using i.i.d. Gaussian rvs $\{u_\ell(t)\}$ with zero mean, and $\{f_\ell(t)\}$ is a sequence of i.i.d. Rayleigh rvs.

We denote the spatial covariance matrix of user j as $\mathcal{H}_j$. The SINR of user j, denoted by $SINR_j$, is given by $S_j/(I_j^{intra} + n_j^2)$, where S_j and I_j^{intra} are the signal power and intra-cell interference received by user j, respectively, and n_j^2 is the noise power at user j. Let $\mathcal{N}$ be the set of co-cell users. Then, the signal power S_j and the intra-cell interference I_j^{intra} are given by

$$S_j = p_j \left(\mathbf{w}_j^H \mathcal{H}_j \mathbf{w}_j\right) \quad \text{and} \quad I_j^{intra} = \sum_{\substack{k \in \mathcal{N} \\ k \neq j}} p_k \left(\mathbf{w}_k^H \mathcal{H}_j \mathbf{w}_k\right),$$

where p_j and $\mathbf{w}_j$ are transmission power and beamforming vector, respectively, for user $j \in \mathcal{N}$.

B. Optimal beamforming algorithms

In this subsection we briefly summarize a *downlink* beamforming algorithm proposed in [17]. This algorithm equalizes the ratios of the achieved SINRs to some target SINRs (called relative SINRs) in a single cell under the assumption that the noise power at each user is constant and is available to the BS. It consists of two phases; in the first phase, the minimum relative SINR among the scheduled users is maximized (Algorithm I). This is equivalent to finding the largest common relative SINR η_c^* under a power budget constraint $||\mathbf{p}||_1 = P_{max}$, where $\mathbf{p}$ is the transmission power vector. Here $||\cdot||_1$ denotes L^1 norm. Let $\mathbf{W} = \{\mathbf{w}_j, j \in \mathcal{U}\}$ be the ensemble of beamforming vectors, where $\mathcal{U}$ is the set of scheduled users and $|\mathcal{U}| = U$, and γ_j denotes the target SINR of user j. A set of users can be scheduled with their respective SINR requirement satisfied if $\eta_c^* \geq 1$, and a set of users that satisfies this condition is called a *feasible set.*

ALGORITHM I: FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{U}$)

STEP 1: Set $n = 0$, $\mathbf{q}^{(0)} = [0, \cdots, 0]^T$, and $\lambda_{max}^{(0)} = \infty$.

STEP 2: While 1, do

- Set $n \leftarrow n+1$. Solve a set of U generalized eigenproblems:

$$\mathbf{w}_j^{(n)} = \arg \max_{||\mathbf{w}_j||=1} \frac{\mathbf{w}_j^H \tilde{\mathcal{H}}_j \mathbf{w}_j}{\mathbf{w}_j^H \mathcal{R}_j(\mathbf{q}^{(n-1)}) \mathbf{w}_j}, \quad \forall j \in \mathcal{U} \tag{4}$$

where $\tilde{\mathcal{H}}_j = \mathcal{H}_j/n_j^2$, and $\mathcal{R}_j(\mathbf{q}^{(n-1)}) = \sum_{k\in\mathcal{U},k\neq j} q_k^{(n-1)}\tilde{\mathcal{H}}_k + I$ with I being the $M \times M$ identity matrix. The solutions to the above generalized eigenproblems are given by the dominant generalized eigenvectors of the matrix pairs $[\tilde{\mathcal{H}}_j,\ \mathcal{R}_j(\mathbf{q}^{(n-1)})]$ for all $j \in \mathcal{U}$.

- Find the largest eigenvalue $\lambda_{max}^{(n)}$ of $\Lambda(\mathbf{W}^{(n)}, P_{max})$ and the corresponding eigenvector $\mathbf{q}_{ext}^{(n)}$ of the form $\mathbf{q}_{ext}^{(n)} = [\mathbf{q}^{(n)};\ 1]$, where

$$\Lambda(\mathbf{W}, P_{max}) = \begin{bmatrix} \mathbf{D}(\mathbf{W})\Psi^T(\mathbf{W}) & \mathbf{D}(\mathbf{W})\sigma \\ \frac{1}{P_{max}}\underline{1}^T\mathbf{D}(\mathbf{W})\Psi^T(\mathbf{W}) & \frac{1}{P_{max}}\underline{1}^T\mathbf{D}(\mathbf{W})\sigma \end{bmatrix},$$

  $\underline{1} = [1\ \cdots\ 1]^T$, $\mathbf{D}(\mathbf{W}) = \text{diag}\big\{\gamma_1/(\mathbf{w}_1^H\mathcal{H}_1\mathbf{w}_1), \cdots, \gamma_U/(\mathbf{w}_U^H\mathcal{H}_U\mathbf{w}_U)\big\}$, $\sigma = [n_1^2, \cdots, n_U^2]^T$, $\mathbf{\Psi}(\mathbf{W}) = [\psi_{ij}, i, j \in \mathcal{U}]$, and the interference caused by user j to user i per unit power $\psi_{ij} = \mathbf{w}_j^H\mathcal{H}_i\mathbf{w}_j$ if $i \neq j$ and $\psi_{ij} = 0$ if $i = j$.
- If $\lambda_{max}^{(n-1)} - \lambda_{max}^{(n)} \leq \varepsilon$, break.

STEP 3: If $\lambda_{max}^{(n)} \leq 1$, *i.e.,* the set of users can be scheduled in the same timeslot, output 1. Otherwise, output 0.

It is shown [17] that the sequence of eigenvalues $\{\lambda_{max}^{(n)}\}$ is monotonically decreasing and converges to the global minimum λ_{max}^*, which is related to the maximum common SINR ratio η_c^* by the relation $\eta_c^* = (\min_{\mathbf{W}} \lambda_{max}(\Lambda(\mathbf{W}, P_{max})))^{-1} = (\lambda_{max}^*)^{-1}$.

The second phase of the algorithm attempts to minimize the total transmission power subject to SINR requirement given a feasible set (Algorithm II).

ALGORITHM II: MINIMIZE_POWER(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j,\ \forall j \in \mathcal{U}$)

STEP 1: Set $n = 0$, and $\mathbf{q}^{(0)} = [0, \cdots, 0]^T$.

STEP 2: While 1, do

- Set $n \leftarrow n + 1$. Solve a set of U generalized eigenproblems defined in (4).
- Compute $\mathbf{q}^{(n)} = (I - \mathbf{D}(\mathbf{W}^{(n)})\mathbf{\Psi}^T(\mathbf{W}^{(n)}))^{-1}\mathbf{D}(\mathbf{W}^{(n)})\underline{1}$.
- If $||\mathbf{q}^{(n-1)}||_1 - ||\mathbf{q}^{(n)}||_1 \leq \varepsilon$, break.

STEP 3: Compute the optimal downlink transmission power vector given by $\mathbf{p}^{(n)} = (I - \mathbf{D}(\mathbf{W}^{(n)})\mathbf{\Psi}(\mathbf{W}^{(n)}))^{-1}\mathbf{D}(\mathbf{W}^{(n)})\underline{1}$.

STEP 4: Output $\mathbf{p}^{(n)}, \mathbf{W}^{(n)}$.

IV. PROPOSED ALGORITHMS

If the user channels are time invariant, *i.e.,* constant, the optimal policy described in Section II-D can be adopted. In other words, each BS may be able to exhaustively search through all possible transmission rate vectors *off-line*, and select the solution to (3) in each timeslot. However, when channels vary with time, this exhaustive search becomes too computationally expensive and

impractical, if not impossible. This is because the number of possible rate vectors is given by

$$c_0 = \sum_{j=1}^{N} \binom{N}{j} V^j = (V+1)^N - 1 \ , \tag{5}$$

which increases exponentially with the number of co-cell users N. Hence, we turn to the problem of designing heuristic algorithms that will perform well and demand much lower computational requirement.

A. First Proposed Algorithm

Although the optimal policy in (3) does not lead to a practical algorithm, it suggests that a good policy should attempt to give higher priority to users with larger queue sizes. This observation is intuitive in the sense that in order to maintain the stability of the system, there should be a balance between (i) maximizing instantaneous system throughput and (ii) keeping the queue sizes from growing without a bound. Therefore, the optimal policy considers the inner product of two vectors, namely $\mathbf{X}(t)$ and $\mathbf{R}^T \mathbf{1}_{I\times 1}$, where the first term is the queue size vector and the latter represents the transmission rate of each user.

A heuristic algorithm that attempts to mimic the behavior of the optimal policy can order the users based on either (i) the transmission rates they can achieve given the current channel state or (ii) queue sizes. The first approach is problematic as the achievable transmission rates of the users depend on the set of scheduled users and it requires searching through all possible feasible rate vectors. Hence, in our first heuristic algorithm we order the users according to their queue sizes and give higher priority to users with a larger queue. More specifically, the algorithm starts with the user with the longest queue, and tries to schedule the users *sequentially* in decreasing order of their queue lengths. Each new user is allocated the highest possible rate such that the SINR requirement is satisfied with the new rate vector. However, when we insert users into the channel sequentially according to their queue lengths, it is possible that a user already scheduled for transmission prevents a number of other users from accessing the channel because the necessary spatial separability cannot be provided. Therefore, in order to improve

the performance of the system further and maintain linear complexity, we will consider several candidate rate vectors and select the one that maximizes (3). More specifically, we will consider P rate vectors out of all possible rate vectors. Clearly, this subset of candidate rate vectors should consist of the rate vectors that are more likely to maximize (3).

We explain how we generate this subset of candidate rate vectors to be considered. Suppose that we form an ordered list of users by decreasing queue size. In order to generate the p-th candidate rate vector, $p = 1, \ldots, P$, of the subset, we first move the p-th user in the list to the head of the list. Then, starting from the head of the list, go down the list sequentially and insert one user at a time using the largest rate that is allowed while maintaining the rates and required SINR values of the previously scheduled users. Note that in some cases, a user may need to be skipped because the user may not be compatible with other users already scheduled. Once the P candidate rate vectors are generated, out of these rate vectors we select the one that maximizes (3). The pseudo-code of this algorithm is provided below.

ALGORITHM III: HEURISTIC_1($x_j(t)$, $\forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{R} = \emptyset$.

STEP 2: For $p = 1$ to P, do

- Form a list $\mathcal{K}$ of users as follows: Insert the user with the p-th largest queue size at the head of the list, and insert the remaining users by decreasing queue size.
- Initialize the rate vector $\mathbf{r} = \underline{0}$ and the set of scheduled users $\mathcal{U} = \emptyset$.
- While $|\mathcal{K}| \neq 0$, do
  - $flag = 0$.
  - Schedule the user at the head of the list, denoted by j^*, $\mathcal{K} = \mathcal{K} \setminus \{j^*\}$, $\mathcal{U} = \mathcal{U} \cup \{j^*\}$ and $\mathcal{V}_1 = \mathcal{V}$
  - While $\mathcal{V}_1 \neq \emptyset$ do
    - $v_m = \max\{\mathcal{V}_1\}$, $r_{j^*} = v_m$, $\mathcal{V}_1 = \mathcal{V}_1 \setminus \{v_m\}$.
    - If FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j$, $\forall j \in \mathcal{U}$) $= 1$, $flag = 1$ where γ_j is the target SINR for $\forall j \in \mathcal{U}$, break;
  - If $flag = 0$, $r_{j^*} = 0$ and $\mathcal{U} = \mathcal{U} \setminus \{j^*\}$.
- $\mathcal{R} = \mathcal{R} \cup \mathbf{r}$.

STEP 3: Among the rate vectors in $\mathcal{R}$, select $\mathbf{r_o} = \arg\max_{\mathbf{r} \in \mathcal{R}} \sum_{j=1}^{N} r_j x_j(t)$.

STEP 4: MINIMIZE_POWER($P_{max}; \mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{U}$)

The complexity of HEURISTIC_1 scheduling algorithm is $O(PNV)$, and hence it increases *linearly* with both the number of candidate rate vectors P and the number of users N.

B. Second Proposed Algorithm

The optimal scheduling policy (3) is not suitable for implementation because its complexity increases exponentially with the number of users as mentioned earlier (see eq. (5)). However, in order to prevent the complexity from increasing with the number of users, in each timeslot an algorithm can first choose a subset of a *fixed* number of users to be considered for scheduling in the timeslot and then carry out the exhaustive search in (3) on the selected users. In other words, the algorithm will mimic the behavior of the optimal policy on a smaller set of users that are selected and do not consider the remaining users for scheduling in the timeslot. To be consistent with the observation that a higher priority should be given to users with larger queue sizes, we select K users with largest queue lengths. Then, an exhaustive search is conducted for all the feasible rate vectors on this subset of users, and the rate vector that maximizes (3) is selected. A pseudo-code of the proposed algorithm is provided below.

ALGORITHM IV: HEURISTIC_2($x_j(t)$, $\forall j \in \mathcal{N}$)

STEP 1: Initialize $\mathcal{R} = \emptyset$.

STEP 2: Select the K users with largest queue lengths. We denote this set of users as $\mathcal{K}$.

STEP 3: Let $\mathcal{S} = \{\mathbf{r} : r_j \in \mathcal{V} \cup \{0\}$ if $j \in \mathcal{K}, r_j = 0$ if $j \notin \mathcal{K}$ $\}$.

STEP 4: For each $\mathbf{r} \in \mathcal{S}$ do

- $\mathcal{U} = \{j : j \in \mathcal{K}, r_j > 0\}$.
- If FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2, \gamma_j$, $\forall j \in \mathcal{U}$) $= 1$, $\mathcal{R} = \mathcal{R} \cup \mathbf{r}$.

STEP 5: Among the rate vectors in $\mathcal{R}$, select $\mathbf{r_o} = \arg\max_{\mathbf{r} \in \mathcal{R}} \sum_{j=1}^{N} r_j x_j(t)$.

STEP 6: MINIMIZE_POWER($P_{max}; \mathcal{H}_j, n_j^2, \gamma_j, \forall j \in \mathcal{U}$)

The complexity of HEURISTIC_2 algorithm is $O(\sum_{k=1}^{K} \binom{K}{k} V^k) = O((V+1)^K)$, and thus the complexity of the algorithm increases exponentially with the size of the subset K.

V. Performance evaluation of a single-cell network

In this section, we evaluate the performance of the proposed heuristic scheduling algorithms in a single cell network, using simulations. We first describe the simulation setup, and then present the numerical results.

A. Simulation setup

1) Wireless channel model: The parameters used in the simulation for modeling wireless channels are listed in Table I, where λ is the wavelength of the carrier electro-magnetic wave.

		P_{max}	10^{15}	PER_{target}	2%
N	10	M	4	r	λ
R	1000	R'	100	L	6
κ	3.5	ρ	0.1	$\mathbf{E}\left[u_\ell(t)\right]$	0
$\mathbf{E}\left[r_\ell(t)^2\right] - \mathbf{E}\left[r_\ell(t)\right]^2$	1.07	$\mathbf{E}\left[f_\ell(t)^2\right]$	1	$\mathbf{E}\left[f_\ell(t)^2\right] - \mathbf{E}\left[f_\ell(t)\right]^2$	1

TABLE I

PARAMETERS USED IN PERFORMANCE EVALUATION OF SCHEDULING ALGORITHMS

2) Transmission rate: We assume that the BS can transmit packets to a user using either a low transmission rate or a high transmission rate. When a low (resp. high) transmission rate is used, one (resp. two) packet is transmitted in a timeslot. The adopted link curves for both the low and high transmission rates are shown in Fig. 2. These link curves give the probability of a packet transmission resulting in error as a function of achieved SINR, *i.e.*, PER.

3) Traffic load: In our experiment, we generate an $N \times 1$ random vector $\hat{a}$. The arrival rate vector $\mathbf{A} = s \cdot \hat{a}$ where s is a parameter used to scale the arrival rate. Traffic load is defined as $||\mathbf{A}||_1 = \sum_{j=1}^{N} \mathbf{A}_j$. Since it is difficult to characterize the throughput region using simulations, instead we observe the average delay experienced by the packets with increasing traffic load. Typically when the system loses stability there is a sharp increase in the average delay at some point (called the *knee*) due to the instability.

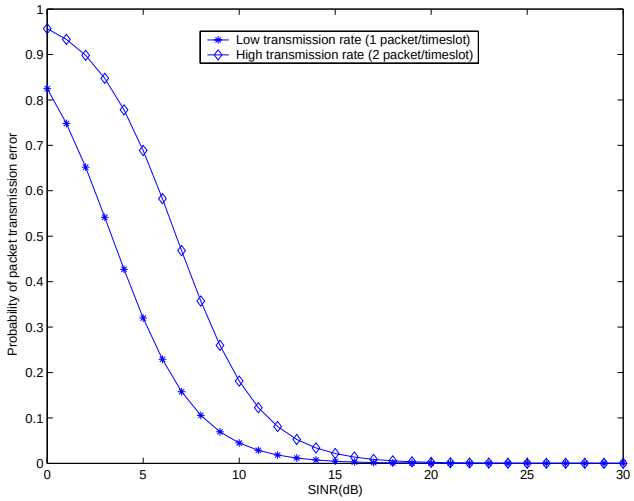


Fig. 2. The linkcurves for low and high transmission rates

B. Numerical results

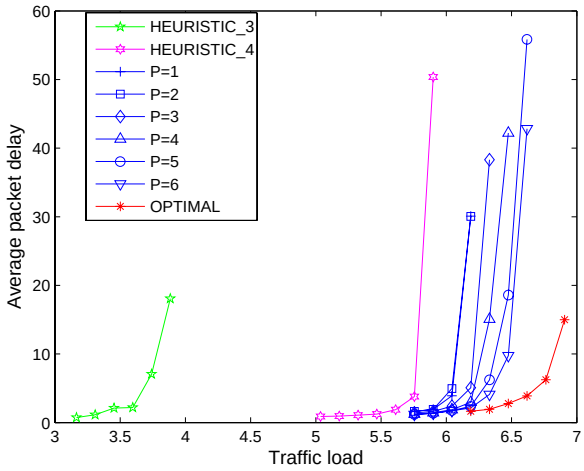


Fig. 3. Average packet delay vs. traffic load for HEURISTIC_1 algorithm for a single cell

In Fig. 3 we show the performance of HEURISTIC_1 algorithm for different values of parameter P. Note that the average delay experienced by packets is similar for $P = 1$ and 2. This is because users are more spatially separable in a single cell network when there is no inter-cell interference, and most of the times first two users with the largest queues can be scheduled with the high transmission rate (2 packets). This can be seen from Fig. 3 since the knee lies right around traffic

load of 6 (larger than 4 = 2 users $\times$ 2 packets). Hence, changing the order of the first two users with the largest queue sizes yields similar rate vectors and does little to increase the number of distinct candidate rate vectors. As a result increasing the number of candidate rate vectors from 1 to 2 does little to reduce the average delay. However, when P is further increased to 3, the maximum stable traffic load (or the knee) increases. This is due to the fact that the subset of candidate rate vectors expands when we first schedule the user with the third largest queue, which otherwise may not be scheduled with high transmission rate (2 packets) when the users are considered in the order of their queue sizes. This can be inferred from the fact that with $P = 1$ or 2 the knee lies slightly to the left of 6 in Fig. 3, indicating that the third user may not always be scheduled with high transmission rate, especially when a fourth user is also scheduled. Hence, this introduces an opportunity to generate a candidate rate vector quite different from those generated by considering only the first two users with high transmission rate. For a similar reason, the maximum stable traffic load increases with $P = 4$, 5, and 6 (compared to $P = 3$). When the value of P is increased beyond 6, the average delay (and hence the maximum traffic load that can be handled without losing stability) does not change much. We suspect that this is due to the fact that the additional candidate rate vectors generated schedule a user with a much smaller queue size and hence the inner products in (3) of these additional candidate rate vectors are smaller than those of previously generated candidate rate vectors that schedule users with larger queue sizes. The maximum stable traffic load of HEURISTIC_1 scheduling algorithm with $P = 1$ (resp. $P = 6$) is approximately 91 percent (resp. 96 percent) of that of the optimal scheduling algorithm as shown in the figure.[2]

For the purpose of comparison, we also evaluate the performance of two other algorithms that do not use the queue length information. The first algorithm, called HEURISTIC_3, assigns a credit c_j to each user, for example, based on the class of service requested by the user, and

[2]Although the plotted average delay of the optimal policy does not increase significantly after the load of 6.7, this is due to the limited simulation run, and increasing the simulation duration results in much larger delays.

keeps track of the throughput $T_j(t)$ of the users. In each timeslot the users are scheduled in increasing order of their ratios of throughput to credit $T_j(t)/c_j$. Since the user with the smallest normalized throughput $T_j(t)/c_j$ is selected for scheduling in each iteration, the algorithm neither takes the queue sizes into account nor attempts to maximize the instant system throughput. In the simulation the credits of the users are set to their arrival rates.

The second algorithm, called HEURISTIC_4, only considers the physical layer separability of the users and attempts to maximize the *instantaneous* throughput of the system in each timeslot without considering the queue sizes [5]. Users are scheduled in the following sequential manner: In each iteration, starting with the highest transmission rate, for each unscheduled user we compute the common relative SINR of the already scheduled users and the user under consideration (using Algorithm I). If there is at least one user that can be scheduled at the selected rate, we select the user that maximizes the common relative SINR. Otherwise, we reduce the transmission rate and repeat the procedure. If no user can be scheduled, we stop. The basic idea behind this algorithm is to schedule a user that is most likely to allow more users to be scheduled in subsequent iterations.

The complexity of HEURISTIC_4 algorithm is $O(MNV)$ when the algorithm is terminated after scheduling at most M users. However, in practice the computational requirement of HEURISTIC_4 algorithm is much higher than that of HEURISTIC_1 algorithm. This is due to the fact that unlike HEURISTIC_4, HEURISTIC_1 algorithm typically does not need to search through all eligible users for generating a candidate vector and searching through 4-5 users (instead of N) is often sufficient. Hence, when N is large, HEURISTIC_1 algorithm requires much lower computation.

The average packet delay under these algorithms is plotted in Fig. 3 as well. As one can see from the plot, the maximum traffic load that can be accommodated under HEURISTIC_3 algorithm is only about 55 percent of that of the optimal scheduling policy, which is considerably smaller than that of HEURISTIC_1 algorithm. The performance of HEURISTIC_4 is better than that of HEURISTIC_3, but is considerably worse than that of HEURISTIC_1 algorithm with $P = 4$ despite

its higher computational complexity. The poor performance of these algorithms is due to the fact that the current queue lengths of the users are not considered for scheduling.

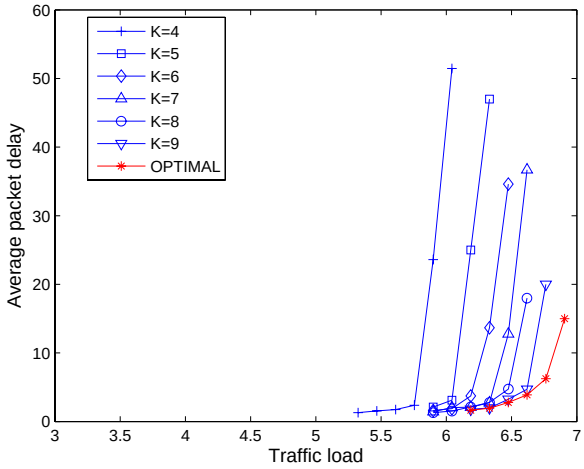


Fig. 4. Average packet delay vs. traffic load for HEURISTIC_2 algorithm for a single cell

We also evaluate the performance of HEURISTIC_2 algorithm, which is plotted in Fig. 4. As expected, the maximum stable traffic load increases with K, the size of the subset of users considered for scheduling in each timeslot. Since there are 10 users in the system, the algorithm with $K = 10$ is the optimal policy because all feasible rate vectors formed by all users are considered. The maximum stable traffic load for $K = 4$ is about 87 percent of that of the optimal scheduling policy in this example.

One thing to keep in mind is that since HEURISTIC_2 algorithm considers all possible rate vectors with a subset of K users, the complexity of the algorithm even for $K = 5$ is higher than that of HEURISTIC_1 algorithm with $P = 6$. Therefore, these simulation results of a single cell network indicate that HEURISTIC_1 algorithm may be preferable to HEURISTIC_2 algorithm as it outperforms HEURISTIC_2 algorithm with lower complexity.

VI. MULTIPLE CELLS NETWORK MODEL

In the previous sections we have considered a simple case where there is only one cell in a network. In practice, in order to improve the spectral efficiency of the cellular system, the

available spectrum is reused in multiple cells, which are called *co-channel* cells. Since the same frequency band is used in more than one cell, these co-channel cells cause inter-cell interference at the users in other cells. Therefore, for a thorough evaluation of the performance of our proposed algorithms, we need to evaluate their performance in the presence of inter-cell interference that reduces the spatial separability of the users as will be shown shortly.

In this section we first describe the network model with multiple cells that we use for our study (subsection VI-A), and then introduce a beamforming algorithm we proposed in [16] for computing the beamforming weights and transmission powers of scheduled users in the presence of *random* inter-cell interference (subsection VI-B).

A. Multi-cell channel model

We consider a network consisting of 7 co-channel cells shown as the shaded cells in Fig. 5. The co-channel cells of a cell can be found by (1) moving x cells along any chain of hexagons, (2) turning 60 degrees counter-clockwise, and (3) moving y cells ([13, p. 28]). A total of $x^2+x\cdot y+y^2$ cells share the available spectrum. We call this pair (x, y) the reuse pattern. The radius of a cell is denoted by R. One BS is located at the center of a cell.

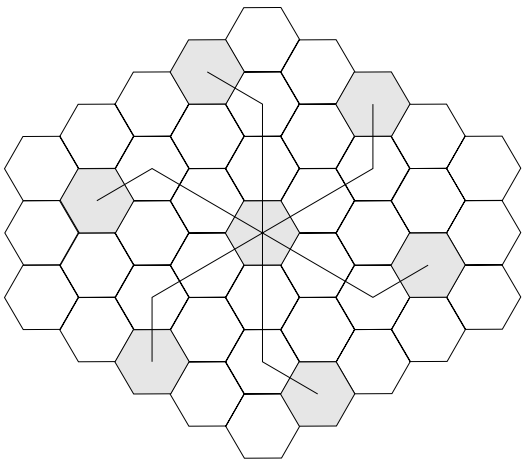


Fig. 5. Co-channel cells ($x = 2$ and $y = 1$).

The channel model we adopt for a user in a cell to each BS (including co-channel cell

BSs) is the same as the channel model used in the single cell scenarios, which is described in subsection III-A. We denote the spatial covariance matrix of user j with respect to BS b as $\mathcal{H}_j^b$ and the BS assigned to user j as b_j. The SINR of user j, denoted by $SINR_j$, is given by $S_j/(I_j^{intra} + I_j^{inter} + n_j^2)$, where S_j, I_j^{intra}, and I_j^{inter} are the signal power, intra-cell interference, and inter-cell interference received by user j, respectively, and n_j^2 is the noise power at user j. Let $\mathcal{U}^\star$ be the set of users in the 7 co-channel cells we consider. For each user j,

$$S_j = p_j \left(\mathbf{w}_j^H \mathcal{H}_j^{b_j} \mathbf{w}_j\right), I_j^{intra} = \sum_{\substack{k \in \mathcal{U}^\star \\ k \neq j, b_k = b_j}} p_k \left(\mathbf{w}_k^H \mathcal{H}_j^{b_k} \mathbf{w}_k\right), \text{ and } I_j^{inter} = \sum_{\substack{k \in \mathcal{U}^\star \\ k \neq j, b_k \neq b_j}} p_k \left(\mathbf{w}_k^H \mathcal{H}_j^{b_k} \mathbf{w}_k\right)$$

where p_j and $\mathbf{w}_j$ are the transmission power and beamforming vector, respectively, for user $j \in \mathcal{U}^\star$.

As mentioned at the beginning of this section, in a multi-cell network the inter-cell interference experienced by a user depends on many factors, including the scheduling and beamforming algorithms used in the co-channel cells. In order to provide satisfactory physical layer QoS to the scheduled users, the beamforming algorithms need to account for the *random* inter-cell interference that will be experienced by scheduled users. In this study we adopt a beamforming algorithm proposed in [16] that is based on a closed-form approximation of PER as a function of SINR to handle the issue of random inter-cell interference experienced by the users. This algorithm is summarized in the following subsection.

B. Beamforming algorithm for multi-cell networks

We observe that most of the link curves (*e.g.*, [8], [12]) can be fitted using a function of the form

$$PER(SINR) = 1/(1 + e^{k(SINR_{dB} - z)}) \tag{6}$$

where $SINR_{dB}$ is the SINR in dB, *i.e.*, $SINR_{dB} = 10\log_{10}(SINR)$, and k and z are two fitting parameters that determine the slope and the position of a link curve, respectively.

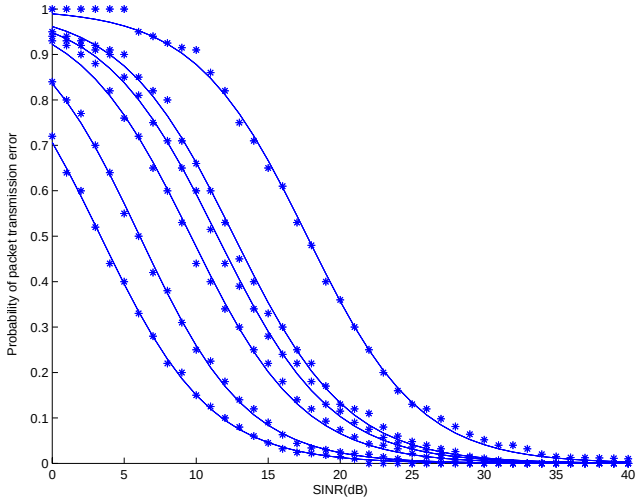


Fig. 6. Link curves of a TDMA system.

The measured/collected data points for different modulation and/or coding schemes of a TDMA system [8] are plotted as '$*$' in Fig. 6, and the fitting curves are shown as solid curves. One can clearly see that these fitting curves match the measured link data very closely. Link curves for systems other than this TDMA system are similar in shape but with different slopes. Several example link curves for a CDMA system are given in [12]. The slope of a link curve reflects the sensitivity of PER to SINR value and is determined by the modulation and/or coding scheme, packet length, etc.

For a reasonable target PER (less than 5-10 percent), we can approximate (6) as follows:

$$\begin{aligned} PER(SINR) &= 1/(1+e^{k(SINR_{dB}-z)}) \\ &\approx e^{-k(SINR_{dB}-z)} = e^{kz}SINR^{-\alpha} \end{aligned} \tag{7}$$

where $\alpha = 10k/\ln 10$.

The following beamforming algorithm is based on the approximation in (7) and attempts to adjust the virtual target PER provided to the beamforming algorithm (which may be different from the desired target PER) in such a way that the achieved PER equals the desired target PER. We refer the interested readers to [16] for details of the algorithm. The algorithm consists of two subalgorithms, FEASIBILITY_M and MINIMIZE_POWER_M. Algorithm FEASIBILITY_M

tests the feasibility of a set of users with their respective transmission rates, and algorithm MINIMIZE_POWER_M minimizes the total transmission power for a feasible rate vector.

ALGORITHM V: FEASIBILITY_M(P_{max}; $\mathcal{H}_j$, n_j^2, $\hat{I}_j^{inter}(t)$, $\overline{\hat{SINR}_j^{-1}}$, $\overline{\hat{SINR}_j^{-\alpha}}$, $\forall j \in \mathcal{U}$)

STEP 1: If FEASIBILITY(P_{max}; $\mathcal{H}_j, n_j^2 + \hat{I}_j^{inter}(t), \gamma_j, \forall j \in \mathcal{U}) = 1$, where γ_j satisfies $e^{k_j z_j} \gamma_j^{-\alpha_j} = \epsilon_j \cdot PER_{target}$, output 1, otherwise output 0.

In the above algorithm the input variable $\hat{I}_j^{inter}(t)$ is the time average of the inter-cell interference at user j, and $\overline{\hat{SINR}_j^{-1}}$ and $\overline{\hat{SINR}_j^{-\alpha}}$ are the time average of $SINR^{-1}$ and $SINR^{-\alpha}$, respectively. The variables k_j, z_j, and α_j denote the parameters of the link curve corresponding to the transmission rate selected for user j. The parameter ϵ_j is defined to be $\overline{\hat{SINR}_j^{-1}}^{\ \alpha_j} / \overline{\hat{SINR}_j^{-\alpha_j}}$ [16].

ALGORITHM VI: MINIMIZE_POWER_M(P_{max}; $\mathcal{H}_j$, n_j^2, $\hat{I}_j^{inter}(t)$, $\overline{\hat{SINR}_j^{-1}}$, $\overline{\hat{SINR}_j^{-\alpha}}$, $\forall j \in \mathcal{U}$)

STEP 1: MINIMIZE_POWER(P_{max}; $\mathcal{H}_j, n_j^2 + \hat{I}_j^{inter}(t), \gamma_j, \forall j \in \mathcal{U}$), where γ_j satisfies $e^{k_j z_j} \gamma_j^{-\alpha_j} = \epsilon_j \cdot PER_{target}$.

These algorithms are modifications of the beamforming algorithms FEASIBILITY and MINIMIZE_POWER described in subsection III-B, and ensure that the average PER equals the desired target PER in a multi-cell network with inter-cell interference.

VII. PERFORMANCE EVALUATION OF A MULTI-CELL NETWORK

For our simulation with multiple cells, we use a reuse pattern of $(x, y) = (1, 1)$. As with the single cell scenarios, we study the average packet delay with varying traffic load only in the center cell under the proposed scheduling algorithms. The BSs in the surrounding cells utilize HEURISTIC_1 algorithm with $P = 1$, and the traffic load is assumed to be constant.

The performance of HEURISTIC_1 algorithm is shown in Fig. 7. We observe that the maximum stable traffic load achieved by the optimal scheduling policy in (3) for this multi-cell network is

about 80 percent of that of a single cell network due to the presence of inter-cell interference, which reduces the spatial separability of the users.

It is worth noting that unlike in the single cell network, the maximum stable traffic load increases from $P = 1$ to $P = 2$. This is due to the inter-cell interference that makes the users less spatially separable. Therefore, switching the order of the two users with largest queue lengths often yields different rate vectors because two users with the largest queue sizes may not be scheduled together any more due to the lack of spatial separability. This increases the subset of candidate rate vectors, and leads to better performance and larger throughput region. Algorithm HEURISTIC_1 with $P = 2$ and $P = 3$ has similar maximum stable traffic load that is about 90 percent of that achieved by the optimal scheduling policy (3), which is slightly lower than in the single cell scenario (91 - 92.5 percent). This performance degradation of HEURISTIC_1 scheduling algorithm in a multi-cell network is due to the fact that the users are less spatially separable in a multi-cell network in the presence of inter-cell interference. This oftentimes prevents the users with large queue sizes from being spatially separable, and as a result they cannot be scheduled together and sequential scheduling of the users based on their queue lengths may result in a rate vector not close to the optimal one selected by (3).

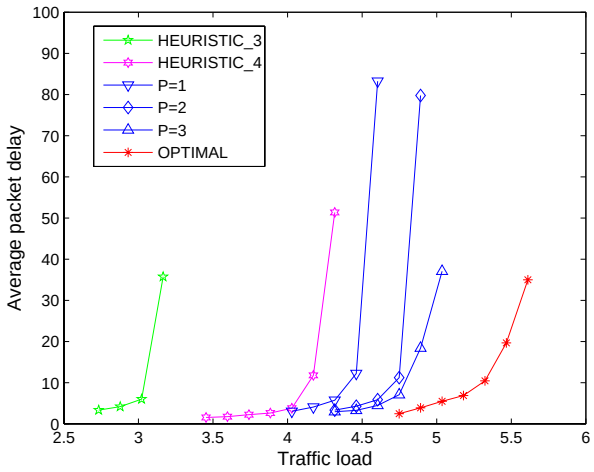


Fig. 7. Average packet delay vs. traffic load for HEURISTIC_1 algorithm for multiple cells

The performance of HEURISTIC_3 and HEURISTIC_4 algorithms is also shown in Fig. 7 for comparison. The maximum stable throughput of HEURISTIC_3 and HEURISTIC_4 is about 55 percent and 78 percent of that of the optimal scheduling policy (3), respectively, in this multi-cell scenario. Thus, our proposed algorithm outperforms the algorithm that attempts to maximize the instantaneous throughput (*i.e.*, HEURISTIC_4) without considering the queue sizes.

The performance of HEURISTIC_2 algorithm with different values of parameter K is displayed in Fig. 8. The maximum stable traffic load increases with K as expected. With $K = 4$, the maximum stable traffic load is about 73 percent of that achieved by optimal scheduling policy, compared to 87 percent for the single cell network. This is again due to the fact that users are less spatially separable with inter-cell interference. Therefore, considering the feasible rate vectors only for a small number of users often selects a rate vector that is not close to the optimal one given by (3).

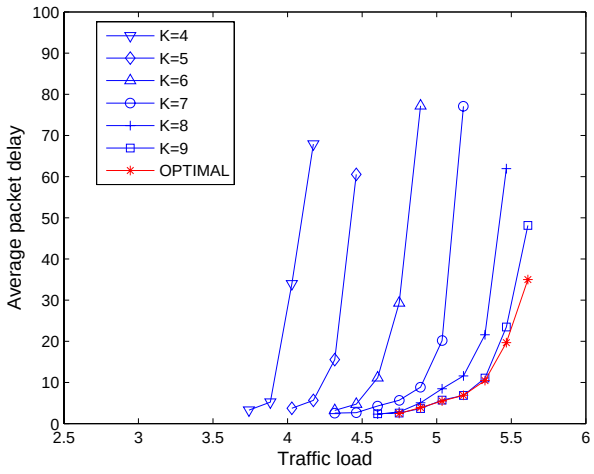


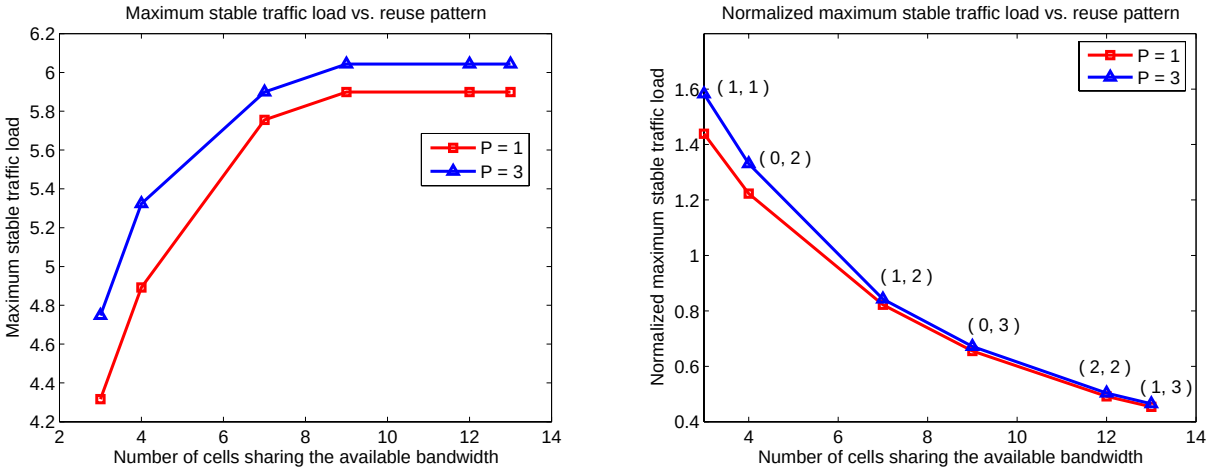
Fig. 8. Average packet delay vs. traffic load for HEURISTIC_2 algorithm for multiple cells

Comparing the figures for single cell and multi-cell networks, we observe that the average packet delay increases more smoothly for the multi-cell network. This is because in a single cell network scenario things are more deterministic due to the absence of inter-cell interference, while in a multi-cell network scenario the presence of time varying inter-cell interference introduces

more stochastic disturbance and randomness to system dynamics.

A. Spectral efficiency under different reuse patterns

In this section we have fixed the reuse pattern at $(x, y) =$ (1,1) and studied the performance of our proposed algorithms against that of the optimal algorithm as well as that of two other heuristic algorithms. In this subsection we are interested in investigating the spectral efficiency of the system under HEURISTIC_1 algorithm with different reuse patterns. Changing the reuse pattern alters the distance between BSs and hence the characteristics of the random inter-cell interference.



(a) (b)

Fig. 9. Plot of (a) maximum stable load and (b) maximum stable load normalized by $x^2 + xy + y^2$ for different reuse patterns

Fig. 9 plots the maximum stable load and the maximum stable load normalized by x^2+xy+y^2 under different reuse patterns. Fig. 9(b) clearly shows that the normalized maximum stable load decreases with the number of cells sharing the spectrum, indicating lower spectral efficiency. In addition, as the distance between BSs increases, the inter-cell interference becomes weaker and the maximum stable load increases and approaches that of a single cell network. Further, the performance gap between $P = 1$ and $P = 3$ narrows with increasing distance between BSs. This is due to the fact that as the inter-cell interference gets weaker, the spatial separability of the users

improves and increasing the number of candidate rate vectors from 1 to 3 does not significantly improve the network performance as mentioned in Section V. This result indicates that in order to improve the spectral efficiency a smaller reuse pattern is preferred under HEURISTIC_1 algorithm when our proposed beamforming algorithm is adopted to handle the random inter-cell interference from co-channel cell BSs.

VIII. DISCUSSION

The use of antenna arrays at the base stations has been proposed to improve the system throughput and to provide quality-of-service (QoS) guarantees to mobile users in wireless networks. In this paper we studied the problem of wireless scheduling with base station antenna arrays with a physical layer constraint of providing certain packet error rate and higher layer QoS guarantees in the form of throughput. An optimal scheduling policy that achieves the throughput region is derived.

We have proposed two heuristic algorithms that attempt to mimic the behavior of the optimal policy with much lower complexity. Simulation results suggest that these algorithms yield significant performance improvement over other algorithms that do not consider queue state for scheduling decisions. The first proposed algorithm is shown to achieve the schedulable region close to the throughput region with linear complexity in the number of candidate vectors and the number of users, whereas the complexity of the optimal policy increases exponentially with the number of users. Furthermore, simulation results indicate that the number of candidate vectors required to enjoy most of the benefits is close to the number of antenna elements at the base stations, which could be orders of magnitude smaller than the number of all feasible rate vectors.

REFERENCES

[1] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a queueing system with asynchronously varying service rates," *Probability in the Engineering and Informational Sciences*, vol. 18, no. 2, pp. 191-217, April 2004.

[2] S. Asmussen, *Applied Probability and Queues*, Wiley 1987

[3] M. Bengtsson, "Jointly optimal downlink beamforming and base station assignment," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, Ut, May 2001

[4] C. Farsakh and J. Nossek, "Spatial covariance based downlink beamforming in an SDMA mobile radio system," *IEEE Transactions on Communications*, vol.46, no.11, pp. 1497-1506, November 1998

[5] I. Koutsopoulos, T. Ren and L. Tassiulas, "The impact of space division multiplexing on resource allocation: a unified approach," *IEEE INFOCOM*, San Francisco, CA, April 2003

[6] S. S. Kulkarni, and C. Rosenberg, "Opportunistic scheduling for wireless systems with multiple interfaces and multiple constraints," *ACM MSWiM*, San Diego, CA, September 2003.

[7] P. R. Kumar and S. P. Meyn, "Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies," *IEEE Transactions of Automatic Control*, vol.41, pp. 4-17, January 1996

[8] K. Leung, P. Driessen, K. Chawla and X. Qiu, "Link adaptation and power control for streaming services in EGPRS wireless networks," *IEEE Journal on Selected Areas in Communications*, vol.19, no.10, pp. 2029-2039, October 2001.

[9] V. A. Malyshev and M. V. Menshikov, "Ergodicity, continuity, and analyticity of countable Markov chains," Transactions of the Moscow Mathematical Society, vol. 39, pp. 1-48, 1979.

[10] M. Neely, E. Modiano and C. Rohrs, "Power and server allocation in a multi-beam satellite with time varying channels," *IEEE INFOCOM*, New York, NY, June, 2002

[11] I. C. Paschalidis, "Large deviations in high speed communication networks," Ph.D Dissertation, MIT LIDS, May 1996.

[12] M. Rajih and S. Sarkar, "Reference link level curves for Qualcomm cdma2000 Revision D R-ESCH," *ftp://ftp.3gpp2.org*, May 2003.

[13] T. Rappaport, *Wireless communications: principles and practice*, Prentice Hall, 2002.

[14] F. Rashid-Farrokhi, K. R. Liu and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE Journal on Selected Areas in Communications*, vol.16, no.10, pp. 1437-1450, October 1998

[15] T. Ren, R. J. La, and L. Tassiulas "Downlink beamforming algorithms with inter-cell interference in cellular networks," *IEEE INFOCOM*, March 2004, Hong Kong, China.

[16] T. Ren and R. J. La, "Downlink beamforming algorithms with inter-cell interference in cellular networks," *IEEE INFOCOM (longer version submitted to IEEE Transactions on Wireless Communications)*, March 2005, Miami, FL. (also available at http://www.ece.umd.edu/~ hyongla/PAPERS/twc04_ren.pdf)

[17] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Transactions on Vehicular Technology*, vol.53, no.1, pp. 18-28, January 2004.

[18] F. Shad, T.D. Todd, V. Kezys and J. Litva, "Dynamic slot allocation (DSA) in indoor SDMA/TDMA using a smart antenna basestation," *IEEE/ACM Transactions on Networking*, vol.9, no.1, pp. 69-81, February 2001.

[19] L. Tassiulas, A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936-1949, December

1992.

[20] L. Tassiulas, "Scheduling and performance limits of networks with constantly changing topology," *IEEE Transactions on Information Theory,* vol.43, no.3, pp. 1067-1073, May 1997

APPENDIX I

PROOF OF PROPOSITION 1

In order to prove the proposition we first show that the throughput region achieved by stationary scheduling policies that consider only the channel state $\mathbf{S}$ is given by the throughput region stated in the proposition. Then, we prove that restricting the scheduling policies to such scheduling policies does not reduce the throughput region.

Define

$$\underline{D}_j := \liminf_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} r_j(\tau)(1 - I_j^e(\tau)) = \liminf_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} d_j(\tau) \tag{8}$$

where $r_j(t)$ is the transmission rate for user j in timeslot t, $d_j(t) = r_j(t)(1 - I_j^e(t))$, $r_j(\mathbf{R})$ denotes the j-th element of the vector $\mathbf{R}^T\mathbf{1}_{I\times 1}$, and $I_j^e(t)$ is an indicator function

$$I_j^e(t) = \begin{cases} 1, & \text{if user } j\text{'s transmission in timeslot } t \text{ is unsuccessful} \\ 0, & \text{otherwise} \end{cases}.$$

These indicator functions $\{I_j^e(t); t = 1, 2, \ldots\}$ are given by a sequence of i.i.d. Bernoulli rvs with $\mathbf{E}\left[I_j^e(t)\right] = PER$. The indicator functions $\{I_j^e(t); t = 1, 2, \ldots\}$, $j = 1, 2, \ldots, J$, for different users are assumed to be mutually independent.

Lemma 1: Consider a single queue j, $j = 1, 2, \ldots, J$. A necessary condition for queue j to be stable is $A_j \leq \underline{D}_j$. Moreover, if the departure process $\{d_j(t); t = 1, 2, \ldots\}$ is given by a finite state, ergodic Markov chain, then a sufficient condition for stability is $A_j < \underline{D}_j$.

Proof: It is well known that with Markovian arrival and departure processes a sufficient condition for the queues to be stable is $A_j < \underline{D}_j$ [10], [11]. Hence, here only prove that $A_j \leq \underline{D}_j$ is a necessary condition.

Suppose that $A_j > \underline{D}_j$. Select $\epsilon > 0$ such that $A_j - \underline{D}_j - 2\epsilon > 0$. We can find a subsequence $\{t_i\}$, where $t_i \to \infty$, such that for all t_i

$$\frac{\sum_{\tau=1}^{t_i} a_j(t_i)}{t_i} \geq A_j - \epsilon \ , \ \frac{\sum_{\tau=1}^{t_i} d_j(\tau)}{t_i} \leq \underline{D}_j + \epsilon \ .$$

Then, it is easy to see that the queue size $x_j(t_i)$ satisfies

$$x_j(t_i) = \sum_{\tau=1}^{t_i} a_j(\tau) - \sum_{\tau=1}^{t_i} d_j(\tau) \geq (A_j - \underline{D}_j - 2\epsilon)t_i \quad \text{for all } t_i \ .$$

Define $\alpha := A_j - \underline{D}_j - 2\epsilon$, and let T_i denote the additional time it takes for the queue size $x_j(t)$ to drop below a threshold value c, starting at the value $x_j(t_i)$ in timeslot t_i. Clearly $T_i \geq (\alpha t_i - c)/v_{max}$, where $v_{max} = \max \ \mathcal{V}$ is the largest transmission rate available. Thus, at time $t_i + T_i$ the fraction of time the queue size exceeds c is lower bounded by $T_i/(t_i + T_i)$, which is greater than or equal to $(\alpha t_i - c)/(\alpha t_i - c + v_{max} t_i)$. Therefore,

$$\begin{aligned} \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}_{[x_j(\tau)>c]} &\geq \lim_{t\to\infty} (\alpha t_i - c)/(\alpha t_i - c + v_{max} t_i) \\ &= \alpha/(\alpha + v_{max}) \ . \end{aligned} \tag{9}$$

Since (9) is true for all $c > 0$,

$$\lim_{c\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}_{[x_j(\tau)>c]} \geq \alpha/(\alpha + v_{max}) > 0 \ ,$$

and the system is not stable. ■

Using a stationary scheduling policy that utilizes only the channel state information leads to a departure process $\{d_j(t); t = 1, 2, \ldots\}$ produced by a Markov chain for all queues j, with an average rate given by the right hand side of the condition in (2). Therefore, Lemma 1 ensures stability when the arrival vector $\mathbf{A}$ satisfies the condition in Proposition 1.

We now proceed to prove that the condition in (2) is a necessary condition even when the above restriction on the scheduling policy is removed. Suppose that all queues can be stabilized with some scheduling policy. Then, from the proof of Lemma 1 a necessary condition for stability is that $A_j \leq \underline{D}_j$ for all users $j \in \{1, 2, \cdots, J\}$, where $\underline{D}_j$ is defined in (8).

Define $1(\mathbf{S},t)$ to be indicator function of the event that channel is in state $\mathbf{S}$ in timeslot t, *i.e.,* $1(\mathbf{S},t)=1$ if $\mathbf{S}(t)=\mathbf{S}$ and 0 otherwise, and

$$1(\mathbf{SR},t)=\begin{cases}1, & \text{if channel is in state } \mathbf{S} \text{ and a rate matrix } \mathbf{R} \text{ is selected in timeslot } t\\ 0, & \text{otherwise}\end{cases} .$$

Fix $\epsilon>0$. There exists sufficiently large $\tilde{t}$ such that

$$\frac{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{S},\tau)}{\tilde{t}}\leq \pi_{\mathbf{S}}+\epsilon \ , \ \underline{D}_j\leq \tfrac{1}{\tilde{t}}\sum_{\tau=1}^{\tilde{t}}d_j(\tau)+\epsilon \ , \quad \text{and}$$
$$\frac{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)(1-I_j^e(\tau))}{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)}\leq 1-PER+\epsilon \quad \text{for all } j=1,2,\ldots,J, \quad .$$

Therefore, for each $j\in\{1,2,\ldots,J\}$ we have

$$\begin{aligned}
A_j=\underline{D}_j &\leq \frac{1}{\tilde{t}}\sum_{\tau=1}^{\tilde{t}}d_j(\tau)+\epsilon\\
&=\frac{1}{\tilde{t}}\sum_{\tau=1}^{\tilde{t}}r_j(\mathbf{R}(\tau))(1-I_j^e(\tau))+\epsilon\\
&=\sum_{\mathbf{S}\in\mathcal{S}}\frac{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{S},\tau)}{\tilde{t}}\frac{1}{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{S},\tau)}\sum_{\tau=1}^{\tilde{t}}1(\mathbf{S},\tau)r_j(\mathbf{R}(\tau))(1-I_j^e(\tau))+\epsilon\\
&\leq\sum_{\mathbf{S}\in\mathcal{S}}(\pi_{\mathbf{S}}+\epsilon)\frac{1}{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{S},\tau)}\sum_{\mathbf{R}\in\mathbf{S}}\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)r_j(\mathbf{R})(1-I_j^e(\tau))+\epsilon\\
&=\sum_{\mathbf{S}\in\mathcal{S}}(\pi_{\mathbf{S}}+\epsilon)\sum_{\mathbf{R}\in\mathbf{S}}\frac{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)r_j(\mathbf{R})(1-I_j^e(\tau))}{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{S},\tau)}+\epsilon\\
&=\sum_{\mathbf{S}\in\mathcal{S}}(\pi_{\mathbf{S}}+\epsilon)\sum_{\mathbf{R}\in\mathbf{S}}\frac{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)}{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{S},\tau)}\frac{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)r_j(\mathbf{R})(1-I_j^e(\tau))}{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)}+\epsilon \ ,
\end{aligned}$$

where $r_j(\mathbf{R})$ is the j-th element of the vector $\mathbf{R}^T\mathbf{1}_{I\times 1}$.

Define

$$C_{\mathbf{SR}}(\tilde{t})=\frac{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)}{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{S},\tau)} \ .$$

Then,

$$\begin{aligned}
A_j &\leq\sum_{\mathbf{S}\in\mathcal{S}}(\pi_{\mathbf{S}}+\epsilon)\sum_{\mathbf{R}\in\mathbf{S}}C_{\mathbf{SR}}(\tilde{t})r_j(\mathbf{R})\frac{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)(1-I_j^e(\tau))}{\sum_{\tau=1}^{\tilde{t}}1(\mathbf{SR},\tau)}+\epsilon\\
&\leq\sum_{\mathbf{S}\in\mathcal{S}}(\pi_{\mathbf{S}}+\epsilon)\sum_{\mathbf{R}\in\mathbf{S}}C_{\mathbf{SR}}(\tilde{t})r_j(\mathbf{R})(1-PER+\epsilon)+\epsilon\\
&\leq\sum_{\mathbf{S}\in\mathcal{S}}\pi_{\mathbf{S}}\sum_{\mathbf{R}\in\mathbf{S}}C_{\mathbf{SR}}(\tilde{t})r_j(\mathbf{R})(1-PER)+\epsilon+\epsilon v_{max}+\epsilon|\mathcal{S}|v_{max}(1-PER+\epsilon)\\
&=\sum_{\mathbf{S}\in\mathcal{S}}\pi_{\mathbf{S}}\sum_{\mathbf{R}\in\mathbf{S}}C_{\mathbf{SR}}(\tilde{t})r_j(\mathbf{R})(1-PER)+\epsilon(1+v_{max}+|\mathcal{S}|v_{max}(1-PER+\epsilon)) \ .
\end{aligned}$$

Since ϵ can be arbitrarily close to zero, this completes the proof.

APPENDIX II
PROOF OF PROPOSITION 2

We follow the same fluid limit approach first shown in [9] and used in [1]. Let

$$\mathbf{Y} = (\mathbf{Y}(t), t \geq 0) := (\{((U_{j1}(t), \ldots, U_{j\mathbf{X}_j(t)}(t)), j \in \mathcal{J}); \mathbf{S}(t)\},\ t \geq 0) \tag{10}$$

denote the stochastic process describing the behavior of the system, where $U_{jk}(t)$ is the waiting time of the k-th packet in queue j and $\mathbf{X}_j(t)$ is the number of backlogged packets in queue j. We define the norm of the state $\mathbf{Y}(t)$ to be $||\mathbf{Y}(t)|| := \sum_{j\in\mathcal{J}}(\mathbf{X}_j(t) + \mathbf{W}_j(t))$, where $\mathbf{W}_j(t)$ is the waiting time of user j's head-of-line packet. Let $\{\mathbf{Y}^{(n)}, n = 1, 2, \ldots\}$ denote a sequence of processes, where $\mathbf{Y}^{(n)}$ has an initial condition that satisfies $\left|\left|\mathbf{Y}^{(n)}(0)\right|\right| = n$. Theorem 4 in [1, p. 200] tells us that the random process $\mathbf{Y}$ in (10) is stable if there exists $\epsilon > 0$ and an integer $T > 0$ such that, for any sequence of processes $\{\mathbf{Y}^{(n)}, n = 1, 2, \ldots\}$, we have

$$\limsup_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\left|\left|\mathbf{Y}^{(n)}(nT)\right|\right|\right] \leq 1 - \epsilon\ . \tag{11}$$

Using this result, we prove that a family of scheduling policies that select a rate matrix $\arg\max_{\mathbf{R}\in\mathbf{S}(t)} \mathbf{X}(t)^T \Xi(\mathbf{R}^T \mathbf{1}_{I\times 1})$ in timeslot t, where $\Xi = \text{diag}(\xi_j, j \in \mathcal{J})$, is throughput optimal for any positive constants ξ_j, $j \in \mathcal{J}$, and the system is stable for any arrival rate vector $\mathbf{A} \in int(\mathcal{A})$. Note that our proposed scheme is a special case with $\xi_j = 1$ for all $j \in \mathcal{J}$.

We first introduce the following processes:

- $F_j(t)$ – the total number of user j's packets that have arrived up to time t, including those in the queue at time $t = 0$.
- $D_j(t)$ – the total number of user j's packets that have been successfully transmitted up to time t.
- $G_{\mathbf{S}}(t)$ – the number of timeslots before time t when the channel state was $\mathbf{S}$.
- $\bar{G}_{\mathbf{SR}}(t)$ – the number of timeslots before time t when the channel state was $\mathbf{S}$ and the rate matrix $\mathbf{R} \in \mathbf{S}$ was selected.

- $\mathbf{U}_j(t) = \inf\{s \leq t | F_j(s) > D_j(t)\}$

For process $\mathbf{Y}^{(n)}$ we assume that packet arrivals begin at time $t = -n$ although the packet service begins at time $t = 0$. For notational simplicity we use

$$\mathbf{Z} := \Big\{((F_j, D_j, \mathbf{X}_j, \mathbf{U}_j, \mathbf{W}_j), j \in \mathcal{J}), ((G_{\mathbf{S}}, \mathbf{S} \in \mathcal{S}), (\bar{G}_{\mathbf{SR}}, \mathbf{S} \in \mathcal{S}; \mathbf{R} \in \mathbf{S}))\Big\} \ .$$

Here the processes $F_j^{(n)}, \mathbf{W}_j^{(n)}$ and $\mathbf{U}_j^{(n)}$ are defined on $[-n, \infty)$ [1].[3] Note that a sample path of random process $\mathbf{Z}$ uniquely defines that of random process $\mathbf{Y}$.

We denote by $\mathbf{z}^{(n)}(t)$ the scaled process $\frac{1}{n}\mathbf{Z}^{(n)}(nt)$. The following lemma tells us that, for any sequence $\{\mathbf{Z}^{(n)}, n = 1, 2, \ldots\}$, there exists a subsequence $\{\mathbf{Z}^{(k)}, k \in \mathcal{K} \subset \mathbb{N}\}$, where $\mathbb{N} = \{1, 2, \ldots\}$, such that the scaled process converges to a fluid process. Due to the space constraint, we omit the proof of Lemmas 2 and 3.

Lemma 2: With probability one, for any sequence of processes $\{\mathbf{Z}^{(n)}, n = 1, 2, \ldots\}$, there exists a subsequence $\{\mathbf{Z}^{(k)}, k \in \mathcal{K} \subset \mathbb{N}\}$ such that, as $k \to \infty$, the following convergence holds for the scaled subsequence $\{\mathbf{z}^{(k)}, k \in \mathcal{K}\}$:

$$(f_j^{(k)}(t), t \geq -1) \Rightarrow (f_j(t), t \geq -1) \ , \ (f_j^{(k)}(t), t \geq 0) \to_{u.o.c} (f_j(t), t \geq 0)$$

$$(d_j^{(k)}(t), t \geq 0) \to_{u.o.c.} (d_j(t), t \geq 0) \ , \ (\mathbf{x}_j^{(k)}(t), t \geq 0) \to_{u.o.c.} (\mathbf{x}_j(t), t \geq 0)$$

$$(g_{\mathbf{S}}^{(k)}(t), t \geq 0) \to_{u.o.c.} (g_{\mathbf{S}}(t), t \geq 0) \ , \ (\bar{g}_{\mathbf{SR}}^{(k)}(t), t \geq 0) \to_{u.o.c.} (\bar{g}_{\mathbf{SR}}(t), t \geq 0)$$

$$(\mathbf{u}_j^{(k)}(t), t \geq 0) \Rightarrow (\mathbf{u}_j(t), t \geq 0) \ , \ (\mathbf{w}_j^{(k)}(t), t \geq 0) \Rightarrow (\mathbf{w}_j(t), t \geq 0)$$

where $\Rightarrow$ denotes convergence at every continuity point of the corresponding limit function, and $u.o.c.$ denotes uniform convergence on compact intervals.

The limit functions $f_j, d_j, \mathbf{x}_j, g_{\mathbf{S}}, \bar{g}_{\mathbf{SR}}, \mathbf{u}_j$, and $\mathbf{w}_j$ satisfy the following properties: (i) f_j are right continuous with left limits (RCLL), non-negative, and non-decreasing in [-1, ∞), (ii) $f_j, d_j, g_{\mathbf{S}}, \bar{g}_{\mathbf{SR}}$ are non-negative, non-decreasing and Lipschitz continuous in [0, ∞), (iii) $\mathbf{x}_j$

[3]Here we extend the discrete time processes to continuous time processes by assuming that the values of the processes are constant over $[t, t + 1)$.

are continuous in [0, ∞), (iv) $\mathbf{u}_j$ are non-decreasing and RCLL in [0, ∞), and (v) $\mathbf{w}_j$ are non-negative and RCLL in [0, ∞).

In addition, the limit function $\mathbf{z} = (f, d, g, \bar{g}, \mathbf{x}, \mathbf{u}, \mathbf{w})$ satisfies the following properties for all $\mathbf{S} \in \mathcal{S}$ and $\mathbf{R} \in \mathbf{S}$:

$$\sum_{j \in \mathcal{J}} f_j(0) \leq 1 \ , \ f_j(t) - f_j(0) = A_j t \ , \ t \geq 0 \ , \ \mathbf{x}_j(t) = f_j(t) - d_j(t) \ , \ t \geq 0.$$

For any interval $[t_1, t_2] \subset [0, \infty)$,

$$d_j(t_2) - d_j(t_1) \leq \sum_{\mathbf{S} \in \mathcal{S}} \sum_{\mathbf{R} \in \mathbf{S}} r_j(\mathbf{R})(\bar{g}_{\mathbf{SR}}(t_2) - \bar{g}_{\mathbf{SR}}(t_1)) \tag{12}$$

with equality holding if $\mathbf{x}_j(t) > 0$ for all $t \in [t_1, t_2]$. Further, if $d_j(t_1) > f_j(0)$ for some $t_1 \in [0, \infty)$, then for all $t \in [t_1, \infty)$,

$$A_j \mathbf{w}_j(t) = \mathbf{x}_j(t) \ . \tag{13}$$

We define regular points to be the points $t \in [0, \infty)$ at which the derivatives of the limit functions $f_j(\cdot), d_j(\cdot), g_{\mathbf{S}}(\cdot), \bar{g}_{\mathbf{SR}}(\cdot)$ and $\mathbf{x}_j(\cdot)$ exist.

Lemma 3: With probability one, limit function $\mathbf{z}$ satisfies the following property: Suppose that

$$\sum_{j \in \mathcal{J}} \xi_j \mathbf{x}_j(t) r_j(\mathbf{R}) < \max_{\mathbf{R}' \in \mathbf{S}} \sum_{j \in \mathcal{J}} \xi_j \mathbf{x}_j(t) r_j(\mathbf{R}')$$

for some $\mathbf{S}$ and $\mathbf{R} \in \mathbf{S}$ at some regular point $t \geq 0$. Then, $\bar{g}'_{\mathbf{SR}}(t) = 0$.

We now introduce the following quadratic Lyapunov function.

$$L(\mathbf{x}) = \frac{1}{2} \sum_{j \in \mathcal{J}} \xi_j \mathbf{x}_j^2$$

for a vector $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$.

Lemma 4: The limit function $\mathbf{z}$ satisfies the following: (i) $L(\mathbf{x}(t)), t \geq 0$, is an absolutely continuous function, and (ii) for any $\delta_1 > 0$, there exists $\delta_2 > 0$ such that, with probability one, at any regular point t, $L(\mathbf{x}(t)) \geq \delta_1$ implies $\frac{d}{dt} L(\mathbf{x}(t)) \leq -\delta_2$.

Proof: For notational simplicity we assume $PER = 0$, *i.e.,* all packet transmissions are successful, in this proof. Let $\mathcal{R} := \cup_{\mathbf{S} \in \mathcal{S}} \mathbf{S}$ be the set of all possible rate matrices from which

the scheduler can choose, $\mathbf{r} := |\mathcal{R}|$, and $\mathbf{s} := |\mathcal{S}| < \infty$. Let Φ denote an $\mathbf{s} \times \mathbf{r}$ stochastic matrix such that $\Phi_{\mathbf{SR}}$ is the probability that rate matrix $\mathbf{R} \in \mathcal{R}$ is selected when the channel state is $\mathbf{S}$ and $\sum_{\mathbf{R}\in\mathcal{R}} \Phi_{\mathbf{SR}} = 1$. We assume that $\Phi_{\mathbf{SR}} = 0$ if $\mathbf{R} \notin \mathbf{S}$. Given a stochastic matrix Φ, let

$$v_j(\Phi) = \sum_{\mathbf{S}\in\mathcal{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} \Phi_{\mathbf{SR}} r_j(\mathbf{R}) \ .$$

At any regular point t,

$$\begin{aligned}\frac{d}{dt}L(\mathbf{x}(t)) &= \sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j(t)(A_j - d_j'(t)) \\ &= \sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j(t)(A_j - v_j(\bar{\Phi})) + \sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j(t) v_j(\bar{\Phi}) - \sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j(t) v_j(\hat{\Phi}(t))\end{aligned}$$

where $\bar{\Phi}$ is a stochastic matrix corresponding to a stationary scheduling policy that satisfies $v(\bar{\Phi}) - \mathbf{A} \geq \varepsilon \cdot \mathbf{1}_{J\times 1}$ for some $\varepsilon > 0$, $\hat{\Phi}_{\mathbf{SR}}(t) = \bar{g}'_{\mathbf{SR}}(t)/\pi_{\mathbf{S}}$, and the last equality follows from

$$d_j'(t) = \sum_{\mathbf{S}\in\mathcal{S}} \sum_{\mathbf{R}\in\mathbf{S}} r_j(\mathbf{R}) \bar{g}'_{\mathbf{SR}}(t) \qquad \text{from (12).}$$

The existence of such $\bar{\Phi}$ is guaranteed since $\mathbf{A} \in int(\mathcal{A})$.

Let $\delta_3 > 0$ be a constant such that $L(\mathbf{x}) \geq \delta_1$ implies $\max_{j\in\mathcal{J}} \mathbf{x}_j \geq \delta_3$. Then,

$$\begin{aligned}\sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j(t)(A_j - v_j(\bar{\Phi})) &\leq -(\min_{j\in\mathcal{J}} \xi_j) \cdot \delta_3 \cdot \min_{j\in\mathcal{J}}(v_j(\bar{\Phi}) - A_j) \\ &\leq -(\min_{j\in\mathcal{J}} \xi_j) \cdot \delta_3 \cdot \varepsilon := -\delta_2 \ .\end{aligned}$$

Now in order to complete the proof it suffices to show

$$\sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j(t) v_j(\hat{\Phi}(t)) \geq \sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j(t) v_j(\bar{\Phi}) \ .$$

Define

$$\begin{aligned}K(\Phi, \mathbf{x}) &:= \sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j v_j(\Phi) \\ &= \sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j \sum_{\mathbf{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} \Phi_{\mathbf{SR}} r_j(\mathbf{R}) \\ &= \sum_{\mathbf{S}} \pi_{\mathbf{S}} \sum_{\mathbf{R}\in\mathbf{S}} \Phi_{\mathbf{SR}} \sum_{j\in\mathcal{J}} \xi_j \mathbf{x}_j r_j(\mathbf{R}) \ . \qquad (14)\end{aligned}$$

From (14) it is easy to see that, for any non-negative vector $\underline{x}$, a stochastic matrix Φ maximizes $K(\Phi, \underline{x})$ if and only if $\Phi_{\mathbf{SR}} = 0$ if $\sum_{j\in\mathcal{J}} \xi_j \underline{x}_j r_j(\mathbf{R}) < \max_{\mathbf{R}'\in\mathbf{S}} \sum_{j\in\mathcal{J}} \xi_j \underline{x}_j r_j(\mathbf{R}')$. This property is satisfied for $\underline{x} = \mathbf{x}(t)$ and $\Phi = \hat{\Phi}(t)$ from Lemma 3, and the lemma follows. ■

An immediate consequence of Lemma 4 is that, for any $\delta > 0$, there exists $T > 0$ such that with probability one, $L(\mathbf{x}(t)) \leq \delta$, $t \geq T$.

We now proceed with the proof of the proposition. Lemmas 2 - 4 state that, for any $\epsilon_1 > 0$ we can find a large enough integer $T > 0$ such that, for any sequence of processes $\{\mathbf{Z}^{(n)}, n = 1, 2, \ldots\}$, there exists a subsequence $\{\mathbf{Z}^{(k)}, k \in \mathcal{K} \subset \mathbb{N}\}$ such that, with probability one, the convergence to the limit function takes place (Lemma 2) and $\sum_{j \in \mathcal{J}} \mathbf{x}_j(T) \leq \epsilon_1$ (Lemma 4). If T is sufficiently large, it follows $d_j(T) = f_j(T) - \mathbf{x}_j(T) \geq f_j(T) - \epsilon_1 > f_j(0)$ for all $j \in \mathcal{J}$. This implies by (13) that $\mathbf{w}_j(T) = \mathbf{x}_j(T)/A_j$ for all $j \in \mathcal{J}$. Since ϵ_1 is small, this in turn implies that

$$\sum_{j \in \mathcal{J}} \mathbf{x}_j(T) + \sum_{j \in \mathcal{J}} \mathbf{w}_j(T) \leq (1 + 1/(\min_j A_j))\epsilon_1 := 1 - \epsilon < 1 \ .$$

Thus, with probability one

$$\limsup_{n \to \infty} \frac{1}{n} \left|\left| \mathbf{Y}^{(n)}(nT) \right|\right| \leq 1 - \epsilon \ , \tag{15}$$

and the proposition follows from (11).